Multi-level shared-weight encoding for abstractive sentence summarization

Phi-3:

Zero-Shot:

The evaluation section of the scientific article reveals that the model's performance improved with up to three levels of input sequence encoding, as evidenced by increased Rouge scores. However, performance declined with four or more encoding levels. Computational costs increased with more encoding levels, but weight-sharing helped mitigate this. The model outperformed baselines on the Gigaword Test Corpus, with significant improvements in Rouge scores at three encoding levels. Human evaluations confirmed the model's summaries were grammatically correct and readable, with informativeness and veracity also improving at higher encoding levels. Inter-rater reliability was high, indicating consistent human evaluations.

Few-Shot:

The study evaluates sentiment analysis in U.S. financial news using language models (LLMs) on a dataset of 965,375 articles. Four models were assessed: OPT, BERT, FinBERT, and Loughran-McDonald. OPT showed the highest accuracy, with regression analyses indicating a strong correlation with next-day stock returns. Advanced LLMs like OPT, FinBERT, and BERT outperformed traditional models and were superior for portfolio management strategies. The study concluded that sophisticated LLMs are crucial for accurate sentiment analysis and investment decision-making in financial markets.

Additionally, the study examined the performance of models with varying levels of input sequence encoding. It found that performance improved with up to three encoding levels but declined with higher levels. Computational costs increased with encoding levels, but weight-sharing across levels mitigated this. The study also reported on the models' performance on the Gigaword and DUC2004 corpora, with MED models outperforming baselines. Human evaluations confirmed the high quality of summaries generated by models with three levels of encoding. The study emphasized the importance of encoding levels and computational considerations in model performance and practical applications.

A Machine Learning Algorithm to Identify Patients with Tibial Shaft Fractures at Risk for Infection After Operative Treatment

Phi-3:

Zero-Shot:

In a study evaluating machine learning (ML) algorithms for predicting infection risk after surgery for tibial shaft fractures (TSFs), 1,822 patients were analyzed. The median age was 38, with 75% being men. The study identified 7 key variables for algorithm development, including fracture classification and patient age.

The ML models' performance varied, with the best-performing models (penalized logistic regression, Bayes point machine, and boosted decision tree) showing AUCs between 0.80 and 0.82. The final model, based on penalized logistic regression, had a calibration slope of 0.94 and intercept of 0.00, with a Brier score of 0.078.

The study also developed an online prediction tool incorporating the final model, allowing for postoperative infection risk assessment. The tool demonstrated the ability to predict infection probabilities for different patient scenarios.

Few-Shot:

A machine learning algorithm was developed to predict the risk of infection after surgery for tibial shaft fractures. The algorithm was based on a model that outperformed others in terms of calibration and predictive performance. The final model, derived from a penalized logistic regression algorithm, was chosen for its superior calibration and similar performance to other models. The model incorporated several patient and fracture characteristics, including the Gustilo-Anderson classification, bone loss, mechanism of injury, multitrauma, AO/OTA classification, age, and fracture location. These factors were identified as the strongest predictors of infection. The algorithm's performance was validated in a separate cohort, maintaining high accuracy. An online prediction tool was created to facilitate the calculation of infection risk for patients undergoing operative treatment for tibial shaft fractures.

A New Methodology for Computing Semantic Relatedness

Phi-3:

Zero-Shot:

The evaluation section of the article presents the accuracy of a proposed method in determining semantic relatedness within the Solar domain. The method's accuracy was assessed by comparing the Pearson correlation between human-computed similarity scores and the similarity scores generated by the proposed method. The results yielded an accuracy of 0.863, indicating a high level of correlation with human judgment.

The study compared its method against three standard datasets: Miller & Charles (M&C), Robenstein & Goodenough (R&G), and the 353 word similarity dataset. The comparison showed that the proposed method outperformed previously defined measures in these datasets.

The results are summarized in Table 6, which lists various measures, including Path length, Edge-Method, Wn&Palmer, Lin, IC(Corpus), Jiang & Conrath's IC(Corpus), Tversky's Feature, Feature-based Approach using Wikipedia, and our proposed Similarity(our approach) method. The table demonstrates that the proposed method achieved the highest accuracy scores across the different measures, with the highest being 0.863 for the Similarity(our approach) method.

Few-Shot:

The study evaluates semantic relatedness in the Solar domain using global ontologies and various similarity measures. It compares human-computed similarities with those proposed by the method, achieving an accuracy of 0.863. The method's performance is assessed against standard datasets and measures, outperforming most of them.

Efficient machine learning on data science languages with parallel

Phi-3:

Zero-Shot:

This section evaluates the proposed solution's performance through experiments. The setup involved an 8-node parallel cluster with specific hardware configurations, using R and C++ for development. Two data sets from the UCI repository were used, with subsets created for analysis. Descriptive statistics and machine learning model computations were benchmarked against Spark and a parallel DBMS, Vertica. Our solution showed competitive performance with Spark, especially when data was already partitioned. Computing descriptive statistics on subsets was efficient, with minimal performance impact across different data sizes. However, for large datasets, a parallel cluster was more effective than a single machine due to overheads associated with data partitioning and transfer. The study highlights the trade-offs between single-machine and parallel processing, suggesting the former is preferable for smaller datasets.

Few-Shot:

This section details the experimental evaluation of the proposed solution, including the setup, data, and comparison with other systems. Experiments were conducted on an 8-node cluster with specifications and using R and C++. Data sets YearPredictionMSD and CreditCard were used, replicated to simulate larger data volumes. Descriptive statistics were computed on subsets of data, and machine learning models were compared with Spark and a parallel DBMS, Vertica. The study also explores the trade-offs between single-machine and parallel processing, showing that smaller data sets can be efficiently handled by a single machine, while larger data sets benefit from parallel processing despite initial partitioning overhead.

Natural language processing in mining unstructured data

Phi-3:

Zero-Shot:

In the evaluation section of the scientific article, the authors present the outcomes of their experiments within the field of computer science. They meticulously detail the results achieved, emphasizing the significance of these findings. The experiments conducted are aimed at testing the efficacy of a particular computational method or algorithm. The results are quantified, with specific figures and statistics highlighting the performance improvements or advancements made. The authors interpret these results, explaining how they contribute to the broader understanding of the subject matter. They also discuss potential implications for future research and practical applications. The summary retains all critical data and numbers, ensuring that the essence of the original text is preserved while adhering to the length constraint.

Few-Shot:

The summary provided is a comprehensive overview of the findings from the evaluation section of a study on sentiment analysis in U.S. financial news using language models (LLMs). The study processed 965,375 articles and evaluated four models: OPT, BERT, FinBERT, and Loughran-McDonald. OPT demonstrated the highest accuracy, followed by BERT and FinBERT, while the traditional Loughran-McDonald model lagged significantly behind. Regression analyses revealed OPT's strong correlation with next-day stock returns, followed by FinBERT and BERT. Factors like model design and training data specificity contribute to these differences. Portfolio management strategies based on sentiment analysis further demonstrated the superior performance of advanced LLMs compared to traditional methods. Long-short OPT strategy outperformed others significantly, indicating the predictive capability of advanced LLMs in forecasting market movements. Overall, the study emphasizes the importance of employing sophisticated language models for accurate sentiment analysis and investment decision-making in financial markets.

The summary also discusses the trends in research in various applications year-wise, highlighting the increase in research in MSR from 2010 to April 2018. It notes that approximately 48 percent of the papers belong to summarization and sentiment analysis, making these topics the most popular in the NLP field. The summary also touches on the increase in research works in mobile analytics from 2012 to 2015, the saturation stage of research in the field of context-based sentiment analysis, and the challenges in norms mining in open-source development communities.

Furthermore, the summary addresses the future scope and challenges in the field of sentiment analysis, norms mining, mobile analytics, and summarization techniques. It highlights the need for more research in context-based sentiment analysis, norms mining in open-source development communities, and the summarization of software artifacts. The summary also discusses the challenges in handling massive mobile user review data, the need for better ways to evaluate summaries, and the challenges in creating abstractive summaries.

Lastly, the summary mentions the need for more research on handling heterogeneous data consisting of code fragments from

Semantic Analysis to Identify Students' Feedback

Phi-3:

Zero-Shot:

The article presents the results of experiments conducted over a semester using data from Blackboard, Facebook, and WhatsApp for four different courses. A lexical analyzer was developed for sentiment analysis, assigning scores between -1 and +1 to words based on their sentiment. The study focused on teaching-related sentiments, assigning positive, negative, and neutral scores to each word. The analysis revealed four key areas: knowledge and understanding, course contents, teaching style, and assessment.

The survey results showed an average score of 76% for knowledge and understanding across all courses, while the automated analysis yielded a slightly lower score of 71%. The teaching method received high scores in surveys (>80%), but the automated analysis showed less than 70%. The discrepancy may be due to students feeling more comfortable expressing their opinions on social media.

For the Computer Architecture course, survey results showed an 80% satisfaction rate for knowledge and understanding, while social media analysis showed a 70% rate. This difference could be attributed to the ongoing discussion about the course throughout the semester.

Regarding assessments, social media data showed an average score of 65%, while survey results indicated a higher average of 70%. Students expressed dissatisfaction with exam results on social media, which may have influenced the lower scores in automated analysis.

Overall, the study suggests that automated sentiment analysis can provide valuable insights into student feedback, complementing traditional survey methods. The findings can help improve curriculum and teaching methodologies.

Few-Shot:

The study analyzed sentiment in educational social media data across four parameters: knowledge and understanding, course content, teaching style, and assessment. Data from Blackboard, Facebook, and WhatsApp over six months were processed using a lexical analyzer and sentiment scoring algorithm. Survey results, collected at the semester's end, were compared with automated analysis. The study found survey results for knowledge and understanding to be higher than automated analysis, with a notable discrepancy in teaching style ratings. Social media sentiment analysis revealed ongoing discussions and dissatisfaction with exam results. The study suggests that automated sentiment analysis can provide more accurate student feedback for curriculum improvement.

Sentiment trading with large language models

Phi-3:
Zero-Shot:

3. Results 3.1. Sentiment Analysis Accuracy in U.S. Financial News In this study, we used LLMs to analyze sentiment in U.S. financial news. We processed a dataset of 965,375 articles from Refinitiv, spanning from January 1, 2010, to June 30, 2023. We used 20% of these articles as a test set. We measured the accuracy of each model in predicting the direction of stock returns based on news sentiment. This accuracy indicates how well the model links the sentiment in financial news with stock returns over a three-day period. We evaluated four models: OPT, BERT, FinBERT and the Loughran-McDonald dictionary. Their performance in sentiment analysis is shown in Table 3. Table 3 Language model performance metrics: accuracy, precision, recall, specificity, and the F1 score for each model. The OPT model is the most accurate, followed closely by BERT and FinBERT. Metric OPT BERT FinBERT Loughran-McDonald Accuracy 0.744 0.725 0.722 0.501 Precision 0.732 0.711 0.708 0.505 Recall 0.781 0.761 0.755 0.513 Specificity 0.711 0.693 0.685 0.522 F1 score 0.754 0.734 0.731 0.508 The results show that the OPT model is the most accurate, followed closely by BERT and FinBERT. The Loughran-McDonald dictionary, a traditional finance text analysis tool, has significantly lower accuracy. This indicates that language models like OPT, BERT, and FinBERT are better at understanding and analysing complex financial news. The precision and recall values further support the superiority of the OPT model; its F1 score, which combines precision and recall, also confirms its effectiveness in sentiment analysis. These findings confirm that language models, particularly OPT, are valuable tools

Few-Shot:

The study examined the accuracy of sentiment analysis in U.S. financial news using language models (LLMs). It processed 965,375 articles and evaluated four models: OPT, BERT, FinBERT, and the Loughran-McDonald dictionary. The OPT model demonstrated the highest accuracy, followed by BERT and FinBERT. The traditional Loughran-McDonald model lagged significantly behind. Regression analyses showed OPT's strong correlation with next-day stock returns, followed by FinBERT and BERT. Factors such as model design and training data specificity contribute to these differences. Portfolio management strategies based on sentiment analysis further demonstrated the superior performance of advanced LLMs compared to traditional methods. The long-short OPT strategy outperformed others significantly, indicating the predictive capability of advanced LLMs in forecasting market movements. Overall, the study emphasizes the importance of employing sophisticated language models for accurate sentiment analysis and investment decision-making in financial markets.

The study also evaluated the predictive power of various LLMs in forecasting stock returns. The regression results showed that the OPT model had the strongest correlation with next-day stock returns, followed by BERT and FinBERT. The Loughran-McDonald dictionary model had the least predictive power. The study also found that the predictive strength of the models increased when both LLMs were used as independent variables in the same regression.

The study further assessed the performance of sentiment-based portfolios. It constructed various sentiment-based portfolios using sentiment scores derived from different language models. The long-short OPT strategy demonstrated the most robust risk-adjusted performance,

as evidenced by its superior Sharpe ratio. The Loughran-McDonald dictionary model-based strategy lagged behind, particularly when compared to the value-weighted market portfolio.

The study concluded that advanced language models, particularly OPT, are valuable tools for analyzing financial news and predicting stock market trends. The study also highlighted the importance of model selection in sentiment-based trading and the superiority of advanced LLMs in guiding investment decisions.

Cyber risk prediction through social media big data analytics and statistical machine learning

Phi-3:

Zero-Shot:

Result and discussion

The evaluation section of the article presents the analysis of data collected from Twitter and the CVE database. The data was cleaned, organized, and analyzed using various methods such as histogram analysis, word cloud and commonality analysis, cluster dendrogram analysis, and pyramid analysis.

The histogram analysis revealed that the cyber risk was occurring through Apache, Apache Struts, Yahoo, and Cisco, with the attack methods using the interactive shell, strutsshell interactive, kitploitstrutsshell, and strutspwn exploit. The word cloud and commonality analysis validated that the keywords used in the Twitter and CVE data collection process were appropriate, as there were discussions about vulnerabilities in both platforms.

The cluster dendrogram analysis showed that the terms Apache, Struts, Shell Command, Interactive, and Exploit were found in adjacent clusters, validating that the cyber risk attack on Apache Struts was using the interactive shell command exploits. The pyramid analysis showed that Apache was the third most frequent unigram in Twitter, compared to the CVE database, validating that Apache had the highest cyber risk frequency.

The data was then used to train and test a prediction model using SML. The model was able to predict the occurrence of cyber risk events with an accuracy rate of 96.73%. The model was also able to identify the most frequently discussed topics related to cyber risk, such as Apache, Struts, and Vulnerability.

Overall, the evaluation section of the article demonstrates the effectiveness of using data analysis methods to identify and predict cyber risk events. The results of this research can be useful for insurance companies to provide early warning to their clients and prevent the spread of risk occurrence claims.

Few-Shot:

The study analyzed sentiment in U.S. financial news using language models (LLMs). It processed 965,375 articles and evaluated four models: OPT, BERT, FinBERT, and Loughran-McDonald. OPT demonstrated the highest accuracy, followed by BERT and FinBERT. The traditional Loughran-McDonald model lagged significantly behind. The study also performed regression analyses, which revealed OPT's strong correlation with next-day stock returns. The study concluded that advanced LLMs outperformed traditional methods in portfolio management strategies based on sentiment analysis. The study emphasized the importance of using sophisticated language models for accurate sentiment analysis and investment decision-making in financial markets.

The Dimensionality of Oral Language

Phi-3:

Zero-Shot:

The study aimed to evaluate language measures using various models. Initially, descriptive statistics and correlations were examined, revealing no severe departures from normality and significant correlations among measures. Two approaches were considered: a bifactor model with a common language factor and language-specific factors for English and Spanish, and construct-specific models based on LARRC. The bifactor model showed the best fit, with acceptable and robust standard errors. However, the construct-specific models also provided valuable insights, particularly regarding the intercorrelations among vocabulary, grammar, and higher-level language constructs. The four-factor model from the second approach, combining English vocabulary and grammar with higher-level language, was chosen due to its good fit and alignment with previous research. This model also highlighted a high correlation between Spanish grammar and higher-level language constructs, suggesting potential for combining these factors.

Few-Shot:

The study utilized descriptive statistics and correlation analyses on a range of language measures from various tests to assess their relationship. The data showed no severe abnormalities, and analyses were adjusted for slight nonnormality. Two approaches were taken to model language data: a bifactor model considering both English and Spanish factors, and a construct-specific model based on previous research. The bifactor model had acceptable fit indices, but faced issues with the orthogonality constraint. The construct-specific model had varying fit across models, with the four-factor model being the most supported due to better discrimination and alignment with previous findings of English language unidimensionality in kindergarten. The study suggests the four-factor model as the best representation of language constructs, considering complexity and parsimony.

Using 2D and 3D Computer Games

Phi-3:

Zero-Shot:

Experiment 1 used a game where participants had to select a square, while Experiment 2 introduced a 3D environment with the Colorblind Maze game. Results were recorded in Tables 2 and 3.

Table 2 showed that in Experiment 1, all confirmed colorblind cases were accurately diagnosed. However, individual features like agility or reflex negatively impacted the diagnostic process, leading to a higher number of false positives.

In Experiment 2, the 3D environment had minimal effect on the diagnostic process, with results similar to the 2D game. However, the development and computing power required for the 3D game made it less practical for screening tests.

The analysis concluded that the simplest game, "select a square," was the most optimal for colorblindness detection due to its speed, simplicity, and low error rate. Despite the potential of all proposed games, the technical requirements and time needed for the 3D game made it less favorable for widespread use.

Few-Shot:

The study conducted by Maciej Laskowski explored the effectiveness of interactive methods for detecting colorblindness. The analysis of experiments using the proposed method revealed it as a promising alternative to existing methods. The first experiment indicated that certain individual features, such as agility or reflex, negatively impacted the diagnostic process, leading to a higher number of false positives. However, all confirmed cases of colorblindness were accurately diagnosed. The transition from a 2D to a 3D environment had minimal effect on the diagnostic process. Despite the potential of the proposed games for colorblindness detection, the simplest game, "select a square", proved to be the most optimal in terms of speed, simplicity, and low error rate. Its low technical requirements and straightforward rules make it accessible to a broader audience.